Reducing Bias in Mixture Estimates: A Computer Program to Bin Alleles

Jeff Bromaghin[1] and Penny Crane[2]

[1]Fisheries and Habitat Conservation, U.S. Fish and Wildlife Service, 1011 E. Tudor Rd., Anchorage, AK 99503
[2]Conservation Genetics Laboratory, Region 7, U.S. Fish and Wildlife Service, 1011 E. Tudor Rd., Anchorage, AK 99503

Introduction

Photo by Ray H. Hansen

Mixed-stock analysis (MSA) using genetic characters is an integral part of research programs estimating stock composition of catches of anadromous salmon and describing migration patterns of anadromous salmon in the high-seas.

Widespread adoption of DNA techniques has lead to increased use of highly polymorphic loci in MSA.

Greater polymorphism can enhance the power of MSA, but is not always beneficial. Researchers often bin alleles to reduce the number of parameters to be estimated in studies using conditional maximum likelihood. Alleles are typically binned based on allele size or frequency, but these methods may result in a loss of information.

We present a program for binning alleles to reduce the number of dimensions in a baseline while simultaneously maintaining the ability of the data to differentiate populations.

Program Method

The data for P populations for a locus contain a certain, fixed, amount of information regarding genetic differences between populations.

Exact tests of homogeneity can be used to test if alleles are similarly distributed across populations, with Monte Carlo simulation to estimate significance, to determine binning strategy.

Example:

	Allele								
Population	1	2	3	4	5	6	7	8	Total
1	45	32	9	5	5	1	2	1	100
2	38	24	15	10	10	2	0	1	100
3	62	11	8	8	8	1	1	1	100
4	12	28	45	5	5	0	3	2	100
5	25	49	15	4	4	2	0	1	100

Alleles 1, 2, 3 are informative;
Alleles 4, 5 are informative but not unique;
Alleles 6, 7, 8 are rare, sampling error may be affecting counts.

Tests of Homogeneity:
Allele 4 and 5: *P*=1.000; pool
Alleles 3 and 8: *P*=0.9188; pool
Alleles 1 and 6: *P*=0.7988; pool
Alleles 3_8 and 7: *P*=0.3384; pool
Largest p-value for all other comparisons=0.0615, may lose information if pool further.

Program Detail

Notation

P = number of populations,
A = number of alleles,
n_{ij} = the number of alleles of type j observed in a sample from population i,
n_i = total number of alleles observed in a sample from population i, and
π_{ij} = true proportion of alleles in population i that are of type j.

Binning Algorithm

1. For any two alleles, the hypothesis of homogeneity across populations is tested:
H_0: $\pi_{ik} = \pi_{im}$ for all i versus the alternative
H_1: $\pi_{ik} \neq \pi_{im}$ for at least two i.

2. The matrix is permuted such that the marginal allele and population frequencies of the entire P by A matrix remain fixed. For each permutation of the matrix, the test statistic, likelihood ratio or Pearson Chi-square, is computed for all allele pairs and its value ψ_p is compared to ψ_o, providing an estimate of the probability distribution of the test statistic. The number of times ψ_p exceeds ψ_o, denoted k, is recorded for each pair of alleles. After the matrix has been permuted K times, the pair of alleles having the largest value of k is identified. The ratio p = k/K is an estimate of the significance of an exact test of the hypothesis that the allele proportions are equal across all populations, and large values of p indicate that the allele proportions are not statistically different among the populations.

3. If p exceeds a specified threshold, p_{max}, the two alleles are binned to form a new allele, and the entire process is repeated with the new P by A data matrix; otherwise the process terminates.

Input file

- •*.bse files for SPAM

Program options

- •test all possible pairs of alleles or alleles adjacent in size only
- •test statistic: likelihood ratio or Pearson Chi square
- •number of permutations for Monte Carlo test of significance
- •*P*-value at which to stop testing
- •random seed

Output

- •*.bse file readable for SPAM
- •log file of which alleles were binned and *P*-value for homogeneity test
- •log file can be used by OptiBin to bin alleles of mixture file for estimation

Availability

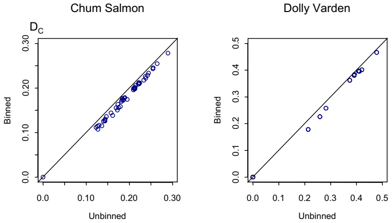
The program OptiBin is available from the website of the U.S. Fish and Wildlife Service, Region 7 Conservation Genetics Laboratory:
http://www.r7.fws.gov/fish/genelab/home.html

Evaluation

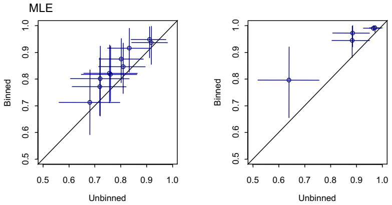
Binning algorithm tested on two data sets, Dolly Varden from western Alaska and chum salmon from the Yukon River.

	Alleles per locus	
	Original	Binned
Chum salmon	9.3	5.4
Dolly Varden	21.3	9

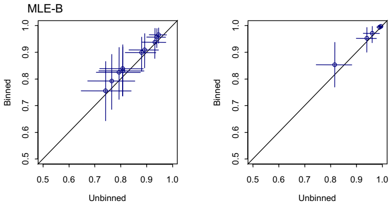
- • Binning reduced the average number of alleles per locus by 50%



- • Binning resulted in a slight reduction in Cavalli-Sforza and Edwards pairwise distances (D_C); change is consistent among all pairs of populations.



- • Binning alleles reduced the bias of the conditional maximum likelihood estimates (MLE) of mixture composition when the maximum likelihood estimator of baseline allele proportions was used.



- • Bias reduction was also apparent, but less substantial, when using a Bayesian estimator of the baseline allele proportions (MLE-B).

Conclusion

This program provides a method to bin alleles that maintains the information content of the data and reduces bias in conditional maximum likelihood estimates.